LANGUAGE-INDEPENDENT TEXT CLASSIFIER BASED ON RECURRENT NEURAL NETWORKS

Vojtech Myska
Doctoral Degree Programme 1st year, FEEC BUT
E-mail: xmyska04@stud.feec.vutbr.cz

Supervised by: Radim Burget
E-mail: burgetrm@feec.vutbr.cz

Abstract: This paper deals with a proposal of language independent text classifiers based on recurrent neural networks. They work at a character level thus they do not require any text preprocessing. The classifiers have been trained and evaluated on a multilingual data set that is privately collected from film review databases. It contains Czech (Slovak), English, German and Spanish language subset. The resulting accuracy of the proposed language independent classifiers base on the recurrent neural networks in polarity sentiment analysis task is 78.55%.

Keywords: sentiment analysis, recurrent neural networks, deep learning

1 INTRODUCTION

Natural language processing (NPL) is one of the fields in which methods based on deep learning have achieved significant progress. One of the NPL tasks is sentiment analysis. This task decides whether the subjective writer's opinion is positive, more positive, neutral, more negative or negative. This paper deals with polarity sentiment analysis task, i.e. only two classes are distinguished - positive and negative.

Recurrent neural networks (RNN) are powerful in extracting patterns of sequence data, such as speech recognition, numerical time series, texts, etc. The proposed models of the classifier are based on RNN variation - LSTM [1].

This paper introduces a novel approach for sentiment analysis task working at the character level. Due to this, the proposed approach does not require any text preprocessing or pre-trained embedding. Therefore, it can be language independent. The introduced approach works at character level instead of word level, or similar higher structures.

2 RELATED WORKS

Text understanding from scratch [3] was published in February 2015 by Xiang Zhang and Yann LeCun. They have demonstrated that knowledge of words, phrases, or some other syntactic or semantic structures associated with a language is not necessary for NPL tasks.

They applied a convolutional neural networks (CNN) models to various text classification tasks. Their proposed model reached up to 95.07% accuracy in the polarity sentiment analysis task.

Character-level convolutional networks [4] was published a few months after the [3]. The paper is an extension of the original work - it contains more experiments of text classification tasks.

Work [2] outperforms the state-of-the-art. In the polarity sentiment analysis task reaches up to 97.84% accuracy.

3 EXPERIMENT

This section describes details of the introduced approach and proposed neural networks models, including a description of the multilingual data set, processing of the text and presentation of achieved results.

3.1 DATA SET

Models proposed in this experiment have been trained and evaluated on the multilingual data set that is privately collected from film review database. This data set is consists of four subsets. Each of them represents Czech (Slovak), English, German or Spanish language and contains 12 000 data samples (text). The experiment deals with polarity sentiment analysis task, thus the subsets have only two numerically balanced categories, i.e. positive and negative.

3.2 TEXT TRANSFORMATION

The input text data has to be converted into a form suitable for neural networks. The text samples are transformed into a 2D matrix. Matrix row dimension is determined by the length of the text, while column by the number of the monitored characters. The text length has been limited to 256 characters. In the experiment 88 characters have been monitored (a-z, 0-9, Czech, Spanish, German, US dollar, euro, pound, space).

Each row of the matrix represents one character of the input. Note that the conversion is sequential. The figure 1 shows an example of converted word "jazz". The algorithm must first obtain the numerical order of the currently processed character in the alphabet. This numerical order determines the column in the matrix. The row is determined by numerical order of the currently processed character in the input text. These two values determine where a value 1 will be written.

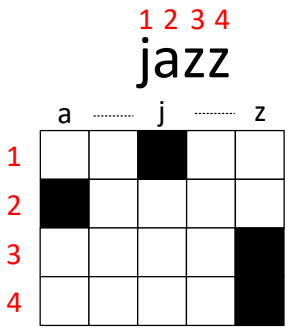

Figure 1: Converted word "jazz".

The proposed neural networks models have classified each text sample as negative if the input data has been in the form above. For this reason, the method of text conversion has been modified. The main idea is to reduce the matrix row dimension by vectors averaging, see figure 2. This modification can be considered very successful because the proposed models have begun to work properly. On the other hands, the method has a disadvantage - the order of characters is not preserved.

The mentioned disadvantage is eliminated by the next method. The key feature is reducing the row dimension while preserving the order of the characters. This has been achieved by concatenating two vectors into one row, see figure 3.

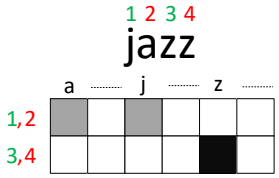

Figure 2: Sub-sampling of matrix row dimension by their averaging.

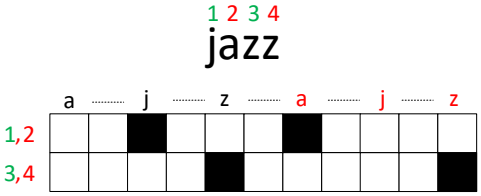

Figure 3: Sub-sampling of matrix row dimension by their concatenating.

3.3 PROPOSED NEURAL NETWORKS MODELS

In this experiment, two neural networks models have been proposed. Both models are based on the same architecture consisting of two main cores, figure 4 shows the architecture. The first core consists of a pair of LSTM layers. The second one is consists of three fully-connected layers. In order to over-fitting, dropout layers are inserted between them.

The first model uses sub-sampling of the matrix row dimension by their averaging. The parameters used in this experiment are listed in table 1. Individual model configurations differ in the number of averaged rows. In this experiment, five models have been evaluated.

Table 2 contains parameters of the model which uses sub-sampling by concatenating rows. Again, five models have been evaluated. These models differ in dropout values.

	LSTM			Dropout	Dense	Softmax	Avg. Pooling
Model	Neurons	Dropout	Recurrent drop.	Drop	Neurons	Neurons	Sub-sampling
1	256	0.18	0.18	0.50	128	2	2
2	256	0.18	0.18	0.50	128	2	3
3	256	0.18	0.18	0.50	128	2	4
4	256	0.18	0.18	0.50	128	2	5
5	256	0.18	0.18	0.40	128	2	6

Table 1: Parameters of the model which uses sub-sampling of matrix row dimension by their averaging.

4 RESULTS

Table 3 shows the effect of reducing the row dimension by averaging rows. Total five models have been tested - they differ in the number of averaging row, see table 1. Model number five has achieved the highest average accuracy of 70.97%. Thus it can be considered as the best. The results show

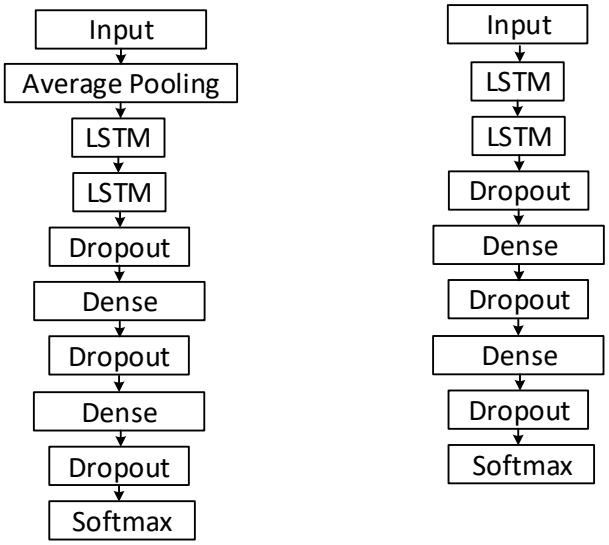

Figure 4: The proposed neural networks models. The first model uses sub-sampling of the matrix row dimension by their averaging. The second one uses sub-sampling by concatenating rows.

	LSTM			Dropout	Dense	Softmax
Model	Neurons	Dropout	Recurrent drop.	Drop	Neurons	Neurons
1	256	0.12	0.12	0.30	256	2
2	256	0.15	0.15	0.35	256	2
3	256	0.15	0.15	0.40	256	2
4	256	0.12	0.12	0.35	256	2
5	256	0.16	0.16	0.45	256	2

Table 2: Parameters of the model which uses sub-sampling by concatenating rows.

that all models have been a little over-fitted. The worst average (64.74%) achieved the first model. Generally, the models have been able to best classify Spanish written texts. The lowest accuracy has been obtained in the classification of German written texts. This can be due to a relatively small data set.

	CZ		EN		DE		ES		Average	
Model	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
1	61.77	60.38	70.3	61.83	61.45	59.92	79.62	76.18	68.28	64.57
2	77.00	71.68	71.02	62.27	78.56	65.70	84.77	79.70	77.83	69.83
3	87.45	75.31	73.72	63.93	79.97	66.84	81.52	76.77	80.66	70.71
4	86.77	74.25	74.52	64.25	79.5	66.83	85.75	78.55	81.63	70.97
5	87.15	73.83	78.81	65.18	80.93	67.33	83.85	76.38	82.6875	70.67

Table 3: Achieved results by the model using row's averaging sub-sampling method.

Table 4 shows results achieved by rows concatenating. Five models have been evaluated. All models reported over-fitting. To thus individual models differ by dropout value, see the table 2. Majority of the models achieved a lower accuracy. The best model achieved only 69.05% accuracy. It is a 1.92% lower accuracy than the best model from the previous results set. As in the previous models classified the best Spanish texts.

	CZ		EN		DE		ES		Average	
Model	Val	Test	val	Test	Val	Test	Val	Test	Val	Test
1	77.75	71.33	80.53	64.06	69.87	63.38	83.62	76.84	77.94	68.90
2	74.47	69.40	57.25	55.13	87.53	66.37	83.58	76.95	75.70	66.96
3	75.48	69.27	70.28	61.67	78.25	65.77	86.25	78.22	77.56	68.73
4	80.98	72.75	78.77	62.77	90.22	65.95	79.80	74.73	82.44	69.05
5	73.7	68.52	64.8	59.57	57.14	59.42	89.10	79.83	71.18	66.83

Table 4: Achieved results by the model using the row's concatenating sub-sampling method.

5 CONCLUSION

This paper shows that the assumption of the possibility of using recurrent neural networks for text classification at character level without sub-sampling of row dimension can not be confirmed yet. The proposed models have not been able to classify text without any matrix modification. In the experiment have been tested models that use rows averaging and concatenating to reduce row dimension.

The accuracy achieved by the first model is from 65.18% (English written texts) up to 78.55% (Spanish written texts). The second model like the first one achieved the lowest accuracy in classification of English written texts (64.04%) and the highest in the classification of Spanish written texts (79.83%). The presented results prove the possibility of classifying texts by recurrent neural networks at the character level without any necessary knowledge of the language, but with reduction row dimension.

REFERENCES

[1] Hochreiter and Schmidhuber: *Long short-term memory, Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.* [cit. 2019-03-12]. Available: `https://arxiv.org/abs/1509.01626`.

[2] Howard and Ruder: *Fine-tuned language models for text classification. CoRR,* vol. abs/1801.06146, 2018. [Online]. [cit. 2019-03-12]. Available: `http://arxiv.org/abs/1801.06146`.

[3] Zhang, Xiang a LeCun, Yann: *Text understanding from scratch. arXiv preprint arXiv:1502.01710, 2015.* [Online]. [cit. 2019-03-12]. Available: `https://arxiv.org/abs/1502.01710`.

[4] Zhang, Xiang a Zhao, Junbo: *Character-level convolutional networks for text classification. Advances in neural information processing systems. 2015, 2015(28), 649-657.* [Online]. [cit. 2019-03-12]. Available: `https://arxiv.org/abs/1509.01626`.